Datamining 4: La regression
Partie 2 A t'on un espoir que ca marche?

M2 STD

November 26, 2014

Problème

On se place désormais dans le cas ou l'on cherche $Y = f(X) + \varepsilon$ avec X dans $\mathbb{R}^d$. On voudrait savoir si la fonction "existe" vraiment ou pas...

- Si $d = 1$ un petit dessin nous en apprend beaucoup....
- Si on suppose que la fonction est linéaire on a le coefficient de corrélation linéaire (qui n'est rien d'autre qu'une analyse du résultat a posteriori en fait)

Le coefficient de Correlation ne marche pas si la fonction est non linéaire !!

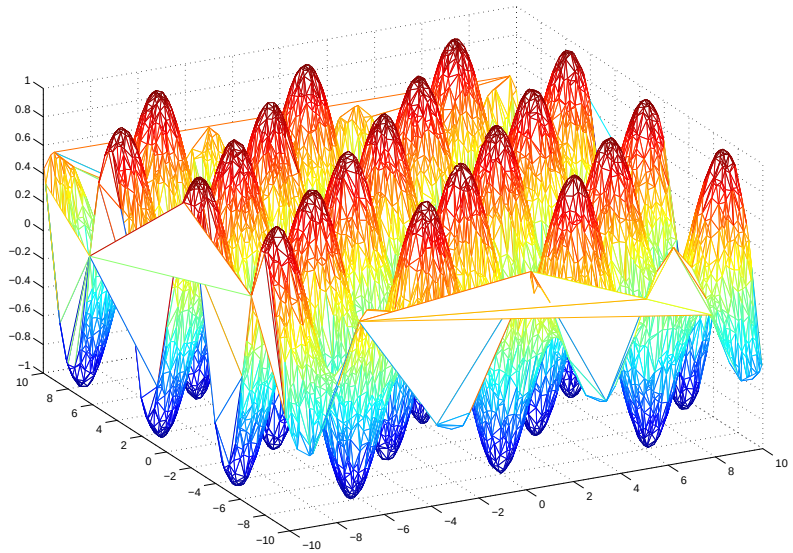

Figure: X uniforme sur $[0,1] \times [0,1]$ et $Y = cos(x) * sin(x)$. Tous les coefficients de corrélations linéaire sont nuls !!!

Rappel

Quand on était dans le cadre de la classification supervisée on avait trouvé une méthode avec les plus proches voisins...
Le problème de la regression est très proche du probleme de la classification (en quelque sorte dual)

- Classification: on cherche $\hat{Y} = f(X)$ avec f discontinue et constante par morceaux
- Regression: on cherche $\hat{f}$ un estimateur de f tel que $Y = f(X) + \varepsilon$ (le plus souvent avec f continu)

Un coefficient de corrélation basé sur les k plus proches voisins

L'idée est toujours la même... Si $k_n \to \infty$ et $k_n/n \to 0$, Notons B_{i,k_n} les boules centrées sur les observations X_i et contenant k_n autres points (des X_j) sont de rayon qui tend vers 0 (on regarde donc les choses localement).
Pour tout i:

- Regardons les Y_j $(j \neq i)$ qui tombent dans la boules, on a $Y_j \sim f(X_i) + \varepsilon_j$, donc $V_i = \hat{V}(Y \cap B_{i,k_n}) \to V(\varepsilon)$
- on peut définir un coefficient de determination comme

$$C(k_n) = 1 - \frac{\frac{1}{n} V_i}{\hat{V}(Y)}$$

et un coefficient de correlation comme la racine du coefficient de determination

Choix du k_n

Comme d'habitude le choix de k_n est primordial... On prendra le k_n qui correspond au "meilleur" estimateur de la regression par la méthode des plus proches voisins (avec bien sur une selection de modèle idoine, p. ex. avec de la cross validation)

Estimateur de la regression par les plus proches voisins

On a principalement deux méthodes:

- La méthode “Naive” a coup de moyennes
- La méthode “Plus éllaborée” a coup de regression locales

Estimateur de regression par plus proches Voisin "naif"

$$\hat{f}(x) = \frac{1}{k_n} \sum_i Y_{i(x)}$$

ou $X_{1(x)}$ est le plus proche voisin de x, $X_{2(x)}$ le second plus proche voisin de x....

Estimateur de regression par plus proches Voisin "regression locale"

$$\hat{f}(x) = \hat{a}_x.x + \hat{b}_x$$

Ou $\hat{a}_x$ et $\hat{b}_x$ sont les coefficients de regression de $(Y_{1(x)}, \ldots, Y_{k_n(x)})$ sur $(X_{1(x)}, \ldots, X_{k_n(x)})$.

dans notre exemple

L'estimateur des pluc proches voisins naif et le regressions locales se comportent de manière similaire... (on choisit alors le plus simple)

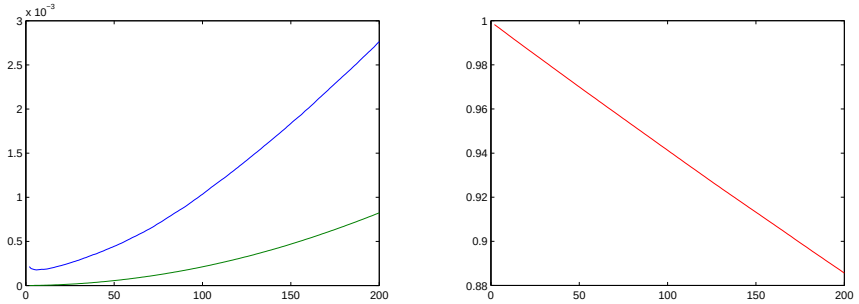

Figure: erreur (gauche, pour l'estimateur naif -bleu- et les regressions locales -vert-) vs coefficient de determination droite